Beaubien, J. M., Holt, R. W., & Hamman, W. R. (1999). Evaluating LOE quality from performance database information. Proceedings of the Tenth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.

EVALUATING LOE QUALITY FROM PERFORMANCE DATABASE INFORMATION

J. Matthew Beaubien Robert W. Holt
George Mason University
Fairfax, VA

Captain William R. Hamman
United Air Lines
Denver, CO

ABSTRACT

Carriers operating under the FAA's Advanced Qualification Program (AQP) are required to maintain databases of crew performance information for use in curriculum refinement and validation. Unlike traditional database management systems, however, crew performance databases must be developed and maintained in such a manner as to allow an assessment of the data's psychometric properties. Using grounded theory and conventional statistical techniques, the authors present a variety of procedures to assist carrier personnel in assessing the usefulness of their crew performance data.

BACKGROUND

Under the AQP, carriers are provided substantial freedom to depart from traditional training methods. However, in exchange for this freedom, the FAA requires carriers to implement data collection strategies and computerized information systems that allow for an independent assessment of the carrier's progress toward achieving their training goals (Federal Aviation Administration, 1998). To facilitate a carrier's decision to enroll in the AQP program, the FAA has sponsored research regarding the development of a model AQP database architecture (Mangold & Neuminster, 1995).

The development and maintenance of crew performance databases, such as the model AQP, represents a fundamental change from traditional data collection, storage, management, and analysis efforts. Unlike traditional Management Information Systems (MIS) or Human Resource Information Systems (HRIS) which contain easily-quantifiable and objective data, crew performance databases are designed for the collection and analysis of performance ratings.

Before performance ratings can be used in conventional statistical analyses, their psychometric properties – sensitivity and reliability – must be established. At the same time, the validity of inferences that can be drawn from such data hinges on the methodological rigor of the overall data collection process (Campbell & Stanley, 1963). In essence, both the type of data collected, as well as their intended use, have engendered a number of issues that were heretofore non-existent in the context of traditional MIS and HRIS systems.

The Components of an AQP Database

Crew performance databases typically include a number of smaller, linked data structures. Common data structures include: a Pilot Characteristics Database (PCDB), a Pilot Performance database (PPDB), a Program Audit Database (PADB), and an Instructor/Evaluator Database (IEDB). Each of these data structures will be discussed in turn.

Pilot Characteristics Database. During the course of a pilot's employment with a given carrier, a wealth of personal information may be collected. This information may include measures of technical proficiency, cognitive ability, personality assessments, attitudes towards CRM, and perceptions of organizational climate. Such data can be extremely valuable, for example as predictors of individual or crew performance during a simulated flight.

Pilot Performance Database. Throughout their careers, pilots are continually re-assessed regarding both CRM and technical proficiency. Typically, technical proficiency is assessed during a maneuvers validation, while CRM proficiency is assessed during line oriented flight training and/or line oriented evaluations. These measures, when combined with evaluations of typical performance collected during a line check, provide a comprehensive portfolio of individual and crew performance over time.

Program Audit Database. Unlike the previous two databases that focus on the individual pilots, a Program Audit Database is specifically designed to

document the training curriculum. Specifically, a program audit database specifies how each terminal objective is to be trained, and where in the curriculum this will occur. Information stored in a program audit database typically includes the interrelationships among terminal objectives, enabling objectives, and knowledges/skills/abilities (KSAs). For example, in a simulated flight, terminal training objectives are linked to specific tasks, which in turn are linked to the KSAs required to perform those tasks.

Instructor/Evaluator Database. Pilot instructors and evaluators are usually former line pilots who currently train and evaluate others. Therefore, it is essential that instructor/evaluators be able to reliably measure performance, and to make meaningful distinctions between crews of different ability levels. The data points stored in an instructor/evaluator database typically include personalized, statistical feedback from rater training programs.

EVALUATION OF LOE QUALITY

For maximum utility, these four databases should systematically linked to one another. For example, because both the crew members' technical proficiency (from the PCDB) as well as the instructor/evaluators' performance ratings (from the IEDB) can influence overall crew performance evaluations in the simulator, these two data files must be integrated.

When properly linked, the information contained in these databases can be used to address hypotheses that are both theoretically meaningful and important to carrier personnel. However, before this can be done, it is essential to develop a conceptual understanding of the data contained within the various data structures. This typically proceeds in a series of steps, starting with the development of a meta-data guide.

Meta-Data Guide

A meta-data guide is a document that describes, in conceptual terms, the different types of data stored in the various databases (Holt, 1998). For example, a meta-data guide might indicate that: the pilot characteristics database contains information regarding personality characteristics and attitudes towards CRM principles; the pilot performance database contains recurrent training data for the past three years; and the program audit database contains the required KSAs for each training maneuver.

The principal value of a meta-data guide is that it provides carrier personnel with an understanding of the types of data which are/are not available for statistical analysis. Quite simply, knowing the type of information available limits the number and type of questions that can be addressed statistically. At the same time, it also suggests the types of information that should be gathered in the future. Developing a meta data guide should be the first step in any serious data collection effort.

Data Codebooks

A data codebook is a document that provides a meticulously detailed description of the information contained within each of the various data structures. For each variable contained within a given database, a codebook typically provides the variable's name, a short description, the measurement scale, missing data codes, and the measure's temporal sequence in the overall data collection process (Cortina et al., 1998).

Data codebooks complement the meta-data guide by providing additional details and insights regarding the overall data collection process. Their primary value lies in the ability to communicate a wealth of information in a very short space. Because AQP-related data is typically gathered at multi-site projects, in which the people who gather the data are different from the people who analyze it, codebooks are essential second step in any data collection effort. Without them, improper data interpretation and needless recalculations will often occur.

Analysis Plans

Once carrier personnel have identified the sources of data available to them, as well as the data's measurement properties, it is now possible to identify a series of substantially meaningful questions that can be addressed by information contained in the database. One such question is the relationship between CRM attitudes and crew behavior. For example, if the researcher knows that the PCDB contains attitudes towards CRM training, and that the PPDB contains ratings of CRM performance, and that both measures are scored on a interval scale, then a product-moment correlation can be used to express the linear relationship between them. In the same manner, a series of systematic research questions can be addressed.

The primary value of the analysis plan is that it organizes the statistical analyses that are desired by

carrier personnel. By addressing research questions in a systematic fashion, the analysis plan allows the data analyst to tell a comprehensive "story" about the data. Typically, this story is more informative and insightful than any question addressed in isolation.

Basic LOE Data Quality Checks

After the various constructs have been selected, their measurement scales identified, and an analytic technique chosen to examine the relationships under consideration, it is incumbent upon the data analyst to evaluate the data's psychometric qualities. Typically, data checks are performed by evaluating the quality of each variable separately, and then using multivariate statistical procedures to understand relationships among combinations of variables.

Missing Data. Missing data is the bane of every researcher. This is due to the fact that missing data reduces one's overall sample size, and thus one's level of statistical power. In many cases, the effects of missing data can be debilitating, especially when listwise deletion procedures are used to form a single sample size for all analyses. At the same time, the presence of missing data can also provide invaluable clues to the astute observer. For example, systematic patterns of missing data can suggest the improper formatting of data collection instruments, poor item wording, unrealistic time constraints, and so forth. Typically, missing data is analyzed for each variable separately with the aid of frequency distributions or histograms. Next, multivariate patterns of missing data can be addressed by searching for patterns of missing data across items (Tabachnick & Fidell, 1996).

Once such patterns have been identified, they should be addressed via consultation with fleet managers and quality assurance personnel. Often, missing data can be easily rectified, such as by including validation codes in database entry forms. Nevertheless, they may require more detailed intervention, such as by re-vamping the overall performance evaluation process.

Range Restriction. The properties of a data set are typically described using measures of central tendency and dispersion. Measures of central tendency, such as the mean, median, and mode, provide information regarding the average value of a given measure. Measures of dispersion, such as the variance and standard deviation, provide information regarding the variability of scores around the mean. Thoroughly examining both types of descriptive statistics is essential to the effectiveness of the entire data-analysis process.

Even though most statistical procedures are based on the assumption that data are normally distributed, this is somewhat unlikely to occur in practice (Murphy & Cleveland, 1995). For example, to the extent that a carrier selects the best qualified candidates and trains them to a carrier-specific benchmark of proficiency, it is unlikely that pilot performance data will be normally distributed. More often than not, most ratings of performance will be clustered at the high end of the rating scale.

Depending on the scale used, this can lead to severe range restriction. For example, if the carrier uses a four-point rating scale to assess crew performance (with a value of three as the minimum acceptable performance rating), virtually all pilots will receive values of three or four. Because there is little variance for any given measure, there can be little covariance between any two measures (Crocker & Algina, 1986). As a result, the carrier may wish to consider using a nine-point scale. Even if the carrier re-scales the rating instrument such that a value of five is the new minimally acceptable performance rating, the scale's increased sensitivity will undoubtedly lead to greater variance, and by extension, greater covariance among measures.

Outlier Analyses. Outlier analyses refer to statistical techniques that are used to determine whether the values for a given variable are plausible. Based on information provided in the codebook, the data analyst can employ basic descriptive statistics such as frequency distributions to identify missing data, out-of-range values, and nonsensical data points. After making the necessary corrections, which often require double-checking the original data-entry forms, means and standard deviations should be re-checked for plausibility as well as range restriction.

After testing for univariate outliers, the presence or absence of multivariate outliers should be systematically investigated (Tabachnick & Fidell, 1996). A multivariate outlier is one in which the value of a single variable, by itself, is not out of the ordinary. However, the combination of values from two or more variables is somewhat unlikely. For example, while it may not be uncommon for a pilot to be in his/her late 20's, or to have five thousand hours of flight time, it would be very uncommon to find a pilot who simultaneously possesses both characteristics.

Level of Analysis. A basic assumption of virtually every statistical procedure is that all observations are independent of one another. However, in aviation settings, in which individuals are nested within crews, this assumption is typically violated. For example, a rating of the captain's performance is likely to be affected by the first officer's performance.

Some variables are clearly measured at the crew level. For example, crews are often required to perform a number of specific tasks per event set. Regardless of which crew member performs the task, both members typically receive the same rating for that item. However, other variables, such as the captain's overall performance rating, are somewhat questionable. For example, is the captain's evaluation measured at the individual or crew level? Or is it a function of both levels?

It is incumbent on the data analyst to establish the level of analysis for all questionable measures. This can be determined via empirical means, such as intra-class correlations or WABA analyses (Bryk & Raudenbush, 1992). At the same time, however, the importance of theory should not be overlooked, as statistical results may suggest that a variable is measured at the group level, even though the results may be due to background experiences that are common to all pilots (Zaccaro, personal communication, 1998). Ignoring levels of analysis issues can lead to inflated sample sizes, spurious statistical significance, biased parameter estimates, and the loss of meaningful variance (Bryk & Raudenbush, 1992).

Quality of the LOE Evaluators

Because LOE performance databases consist largely of crew performance ratings, it is essential that all evaluators receive some form of calibration training. Otherwise, differences among evaluators will lead to error variance that can mask true empirical relationships.

In recent years, we have successfully employed inter-rater reliability (IRR) training programs to assist instructor/evaluators from numerous carriers in understanding their strength and weaknesses as evaluators (Holt et al., 1996). IRR training is conducted in a group session, during which a group of pilot evaluators observe a videotape of crew performance segments, make independent ratings of each segment, and then discuss the reasons for their differences. During the course of training, which is facilitated by subject matter experts, the group comes to some form of consensus, such that when they go back to making their evaluations of crew performance, they will be doing so with a common frame of reference.

Under certain operational conditions, selected IRR analyses can even be applied to the data contained in LOE performance databases. First, instructor/evaluators should be matched with pilot crews in a random fashion. Second, each instructor/evaluator typically should evaluate a large, representative sample of crews. If these conditions exist, two IRR components can be evaluated. First, data averages can be used to assess systematic differences in mean performance ratings among the population of instructor/evaluators. Second, the consistency of each rater's profile of judgments (across items) for the sample with the profile for the entire group can also be assessed.

Once known differences among instructor/evaluators have been identified, they can then be targeted for future training. At the same time, they can also be statistically controlled when evaluating crew performance. This can be achieved in two ways. First, the crew performance ratings made by each instructor/evaluator can be converted into z-scores, and all subsequent analyses can be based on these transformed scores. Second, differences among instructor/evaluators can be controlled by selecting a subset of LOE performance data (e.g., ratings made by only those instructor/evaluators who conform to the rest of the group) in subsequent statistical analyses.

Quality of LOE Content and Process

Once the basic data quality checks have been performed, and the characteristics of the instructor/evaluators assessed, the data can be used to address issues regarding both the LOE content and the LOE rating process. For example, factor analyses can be performed to assess the dimensionality of the LOE measurement scales. In addition, path analyses can be performed to compare the instructor/evaluator's rating process to the carrier's standard operating procedure (SOP) for evaluating crew performance in the simulator. Both techniques will be discussed in turn.

Factor Analyses. The class of multivariate statistical techniques known as factor analysis is used to assess the underlying relationships among a set of variables. For example, if all of the items on a

measure of attitudes towards CRM are hypothesized to be unidimensional, this hypothesis can be tested empirically. If all of the items "cluster" as a single factor, then the data analyst can be confident in the fact that the empirical measure corresponds to the theoretical item structure. However, if the items "cluster" as a number of separate factors, then the researcher should be wary when interpreting the results of overall scale scores in subsequent analyses.

If the data set is novel, exploratory techniques should be probably be used. However if the data set is based on well-established measurement scales, confirmatory procedures may be in order (Byrne, 1998). The principal value of factor analytic techniques is that they allow carrier personnel to empirically test the dimensionality of their rating instruments prior to their use in subsequent analyses.

Path Analyses. Most performance evaluations, such as those collected in an LOE, are presumed to have some type of causal structure. For example, one major carrier requires instructor/evaluators to make three evaluations during any given event set. First, crew members are evaluated on a number of behavioral markers which they are expected to perform during a given phase of flight. These behavioral markers are then used to make crew-level evaluations of CRM and technical proficiency. Finally, the CRM and technical performance ratings are used to make overall evaluations of each crew-members' performance.

The efficacy of this causal chain can be tested empirically using path analytic (Cohen & Cohen, 1983) or structural equation modeling techniques (Byrne, 1998). At times, the data may not be consistent with the hypothesized model. For example, rather than being indirectly linked to overall crew evaluations (e.g., via CRM and technical ratings), behavioral markers may exhibit direct relationships with overall crewmember evaluations. This would indicate that the instructor/evaluators were not making their ratings as per the carrier's SOP. Based on these results, "drill-down" analyses could be performed to localize the root cause of the problem.

A Real-World Illustration

Many of the recommendations suggested in this paper have come about through direct experience gained while collaborating in a three-year research project with a major carrier. Following these recommendations can save much time and effort in data analysis. What follows is a synopsis of a large-scale project that involved the analysis of data from two separate LOEs (Beaubien, Holt, & Hamman, 1999). The study's purpose was to evaluate the rating processes used by instructor/evaluators in the LOE environment.

Data for this analysis was gathered from the pilot performance database (PPDB). Both LOEs contained six separate event sets. Each event set required three separate ratings: behavioral markers, crew-level ratings of CRM and technical performance, and individual ratings of pilot- and second-in-command performance.

Unfortunately, data codebooks were not provided to the research team. Therefore, we had to learn the database structure on our own, and convert it from reduced normal format into a format amenable to statistical analysis. Because of we lacked codebooks, missing data values were not immediately recognized. Several analyses needed to be re-run; some of which changed the fundamental outcome of the analyses. At other times, the incorrect merging of database tables resulted in an artificial "explosion" in sample size. While these setbacks were not catastrophic, they did hamper progress for several weeks until the underlying causes could be identified. Once such issues were resolved, however, we developed a rudimentary data codebook, and relied on an analysis plan that was developed in collaboration with carrier personnel.

At the beginning of the data-analysis process, we began by examining the descriptive statistics. As all variables exhibited severe range restriction, we expected that item covariances would be attenuated. Next, we assessed the level of analysis for all questionable variables. For example, we performed intra-class correlations for ratings of pilot- and second-in-command performance. These analyses suggested that the pilot- and second-in-command ratings contained a sizable "crew-level" component. As all variables were presumed to be measured at the crew level, factor and path analyses were performed without fear of violating the statistical independence assumption.

Next, we performed factor analyses to better understand the content of the measurement scales. In this case, the factor structure of the behavioral markers was virtually uninterpretable. As expected, range restriction suppressed meaningful covariances among items, thereby ensuring that the rating process did not

coincide with the carrier's standard operating procedure.

Armed with this information, the following year's LOE was re-written such that the behavioral markers were both more generalized and fewer in number. The intention was to reduce the instructor/evaluator's cognitive workload in the simulator. Even though the original four-point rating scale was not changed, the data from the second LOE was markedly different from the first.

Specifically, the factor structures among the scale items were much more interpretable. For example, in the second LOE, the behavioral markers formed into six unique clusters (by event set). In addition, the path analyses suggest that the instructor/evaluators were using the rating process as designed. For example, in the second LOE, the magnitude of several paths increased substantially. Because this is an iterative process, feedback from the second LOE is currently being used to make revisions for the third LOE, which is currently being developed.

DISCUSSION

The collection and analysis of crew performance data is part of an iterative process that is focused on improving the training and evaluation of pilot crews, with the ultimate goal being improved line safety. No matter how sophisticated the analysis techniques employed by the data analyst, the results must eventually be interpreted and used by carrier personnel, such as fleet managers and quality assurance monitors. Therefore, it is incumbent on the data analyst to present the data in a format that fleet personnel can readily understand.

Based on our experience with two air carriers, we have found that charts/graphs with one or two short paragraphs of interpretive text are often helpful for summarizing large amounts of data. After all, most fleet managers are former pilots, and pilots tend to be a very visually-oriented sample. We have also found that common presentation formats are helpful. Quite simply, once you find a particular presentation format that fleet personnel find useful, stick with it. It must be remembered that fleet managers have a limited amount of time to devote to any given problem. Unless these reports are easily interpretable, fleet managers may be inclined to pay them little attention.

For statistical feedback to be of use to fleet personnel, it must also be provided on a timely basis. For example, fleet managers may wish to have certain analyses performed each month. Therefore, automation is key. To the extent that certain analyses and reports can be automated, the data-analyst's job becomes that much easier.

Finally, these analyses/reports cannot be performed by a single person working in isolation. In practice, LOE performance databases are complicated, dynamic structures. Like other aspects of organizational reality, they must adapt to changing regulatory conditions, changes in training protocol, and so forth. Therefore, a long-standing, multi-disciplinary group should be established that can tackle the iterative process of data analysis, interpretation, and follow-up. We believe that such groups should be composed of no less than four people: a fleet manager (the end-user), a database specialist, a data analyst, and an individual from training/quality assurance.

REFERENCES

Beaubien, J. M., Holt, R. W., & Hamman, W. R. (1999). An Evaluation of the Rating Process used by Instructor/Evaluators in a Line-Operational Simulation: Preliminary Evidence of Internal Structure Validity. Technical Report #98-002. George Mason University: Fairfax, VA.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS, and SIMPLIS. Mahwah, NJ: Earlbaum.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2^{nd} edition). Hillsdale, NJ: Earlbaum.

Cortina, J. M., Beaubien, J. M., & Holt, R. W. (1998). The Use of Relational Databases in Large Scale, Multi-Site Research Projects: Mitigating the Impact of Data Errors. Technical Report #98-001. George Mason University, Fairfax, VA.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehardt, & Winston.

Federal Aviation Administration. (1998). SFAR 58, Advanced Qualification Program. Available: http://www.faa.gov/avr/AFS/Sfar58.rtf.

Holt, R. W. (1998). Meta-data guide. Unpublished manuscript.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1996). The application of psychometrics to the calibration of air carrier evaluators. Unpublished manuscript.

Mangold, S., & Neuminster, D. (1995). CRM in the model AQP: A preview. In R. S. Jensen and L. A. Rakovan (Eds.)., Proceedings of the Eighth International Symposium on Aviation Psychology (pp. 556-561). Columbus, OH: The Ohio State University Press.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational, and goal-based perspectives. Thousand Oaks, CA: Sage.

Tabachnick, B. G., & Fidell, L. S. (1996). Understanding multivariate statistics (3rd edition). New York: Harper & Row.